第五章 大数定律和中心极限定理

大数定律

中心极限定理

内容1：设$X_1,\ldots,X_n\ldots$是一列随机变量，$EX_i=\mu,$

则在一定的条件下

随机变量序列$Y_n=\dfrac{X_1+\ldots+X_n}{n}$收敛到$\mu.$

问题：（1）一定的条件是什么？

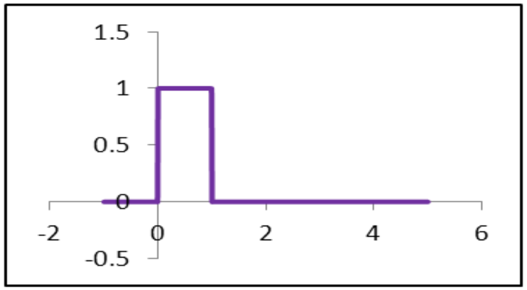
（2）随机变量序列Y_n收敛到μ的定义？

——大数定律

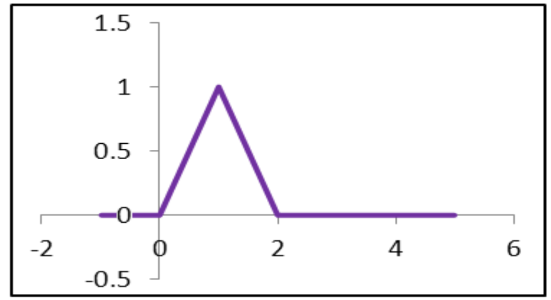
内容2：n个独立同分布随机变量$X_i, i=1,2,\ldots,n$, $\sum_{i=1}^{n} X_i$的分布

例如：$X_i \sim U(0,1), i=1,2,3,4,\ldots,100.$ 则

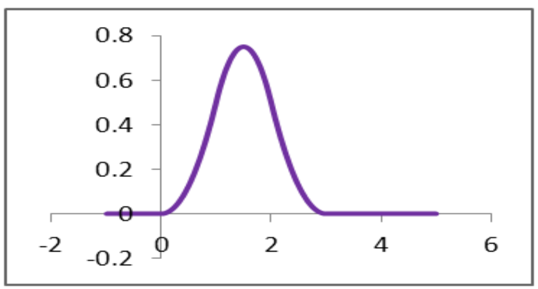
X_1



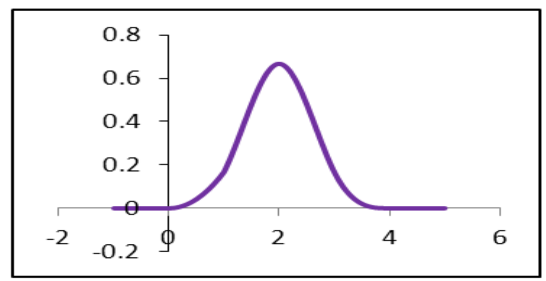
X_1+X_2



$X_1+X_2+X_3$



$X_1+X_2+X_3+X_4$



问题：$X_1+\ldots+X_{100}$ 服从的分布？

——中心极限定理

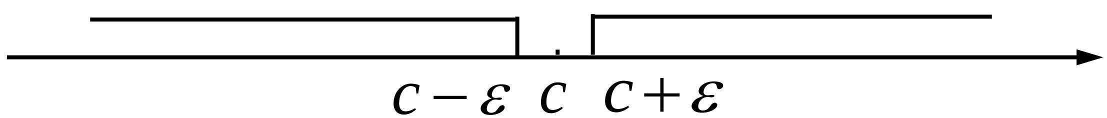
5.1 大数定律

(一) 依概率收敛

随机变量序列$Y_1, Y_2, Y_3, \cdots$，若存在某常数c，使得$\forall \varepsilon > 0$，均有：$\lim\limits_{n \to +\infty} P\left\{\left|Y_n - c\right| \geq \varepsilon\right\} = 0$，

则称$\left\{Y_n, n \geq 1\right\}$依概率收敛于常数$c$，

记为：$Y_n \xrightarrow{P} c$，当$n \to +\infty$时.



这种收敛性是在概率意义下的一种收敛，而不是数学意义上的一般收敛.

例1.1 抛一枚均匀硬币n次，Y_n表示正面出现的频率. n=1, 2 ⋯, 可以证明 $Y_n \xrightarrow{P} 0.5$, 当$n \to +\infty$时. **试说明这种依概率收敛性,**

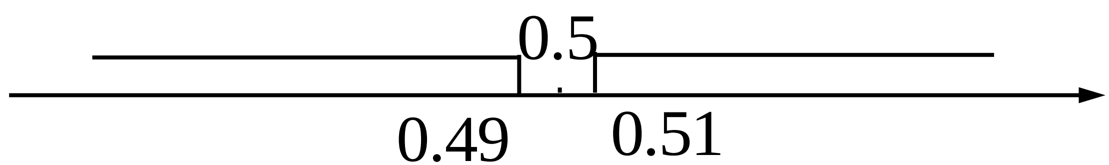
解：这就意味着，$\forall \varepsilon > 0$, 有：$\lim\limits_{n\to+\infty} P\left\{\left|Y_n - 0.5\right| \geq \varepsilon\right\} = 0$,

等价地, $\lim\limits_{n\to+\infty} P\left\{\left|Y_n - 0.5\right| < \varepsilon\right\} = 1$.

例如：$\varepsilon = 0.01$, 有 $\lim\limits_{n\to+\infty} P\left\{\left|Y_n - 0.5\right| < 0.01\right\} = 1$.

计算得：当$n > 6765$时，$P\left\{\left|Y_n - 0.5\right| < 0.01\right\} > 0.90$.

当$n > 9604$时，$P\left\{\left|Y_n - 0.5\right| < 0.01\right\} > 0.95$.



思考题1. 抛硬币7000次，则 $\{0.49 < Y_{7000} < 0.51\}$ 一定发生？

答：不一定. 可能发生，也可能不发生，发生的可能性非常大，概率超过90%.

思考题2. 抛硬币1万次，则事件$\{Y_{10000} \le 0.1\}$会发生吗？说明理由.

答：当$n > 9604$时，$P\{|Y_n - 0.5| < 0.01\} > 0.95. \Rightarrow$

$$P\{Y_n \le 0.49\} = \frac{1}{2}[1 - P\{|Y_n - 0.5| < 0.01\}] \le 0.025.$$

这是一个小概率事件，根据实际推断原理认为

如果只抛1万次，事件 $\{Y_{10000} \le 0.1\}$ 不会发生.

性质：若$X_n \xrightarrow{P} a, Y_n \xrightarrow{P} b, g$在$(a,b)$连续，

则 $g(X_n, Y_n) \xrightarrow{P} g(a,b)$.

如：当$n \to \infty$时，$X_n + Y_n \xrightarrow{P} a+b$,

$X_n \times Y_n \xrightarrow{P} a \times b$, $X_n / Y_n \xrightarrow{P} a/b \ (b \neq 0)$.

特别地，若$X_n \xrightarrow{P} a$, $f(x)$在点a连续，则

$f(X_n) \xrightarrow{P} f(a)$, 当$n \to \infty$时.

例1.2：设$X_n \sim N(0,\frac{1}{n}), n=1,2,\ldots,$

则当$n\to+\infty$时, $X_n \xrightarrow{P} 0$.

证明: 对任意$\varepsilon>0$,

$$P(|X_n-0|\geq\varepsilon)=P(X_n\geq\varepsilon)+P(X_n\leq-\varepsilon)$$

$$=1-\Phi(\frac{\varepsilon-0}{\sqrt{1/n}})+\Phi(\frac{-\varepsilon-0}{\sqrt{1/n}})$$

$$=2[1-\Phi(\varepsilon\sqrt{n})]\to 0,\quad 当n\to+\infty时.$$

(二) 马尔可夫不等式和切比雪夫不等式

定理5.1.1 (马尔可夫不等式):

设随机变量Y的k阶矩存在$(k \geq 1)$,

则对于任意$\varepsilon > 0$, 都有: $P\left\{|Y| \geq \varepsilon\right\} \leq \dfrac{E(|Y|^k)}{\varepsilon^k}$成立;

定理的等价形式为: $P\left\{|Y| < \varepsilon\right\} \geq 1 - \dfrac{E(|Y|^k)}{\varepsilon^k}$.

特别地，当Y为取非负值的随机变量时，

则有 $$P\left\{Y \geq \varepsilon\right\} \leq \frac{E(Y^k)}{\varepsilon^k}$$

证明： 对于任意$\varepsilon > 0$，令$Z = \begin{cases} \varepsilon, & \text{当} |Y| \geq \varepsilon \text{时}; \\ 0, & \text{当} |Y| < \varepsilon \text{时}. \end{cases}$

则$Z \leq |Y|$, $\quad \Rightarrow Z^k \leq |Y|^k$,

$$\Rightarrow E(Z^k) \leq E(|Y|^k).$$

而据Z的定义, 知

$$E(Z^k) = \varepsilon^k P\{|Y| \geq \varepsilon\},$$

所以 $P\{|Y| \geq \varepsilon\} = \dfrac{E(Z^k)}{\varepsilon^k} \leq \dfrac{E(|Y|^k)}{\varepsilon^k}$.

切比雪夫不等式：设X的方差$Var(X)$存在，则

对于任意$\varepsilon>0$,都有：$P\{|X-E(X)|\geq\varepsilon\}\leq\dfrac{Var(X)}{\varepsilon^2}$.

等价为：$P\{|X-E(X)|<\varepsilon\}\geq 1-\dfrac{Var(X)}{\varepsilon^2}$.

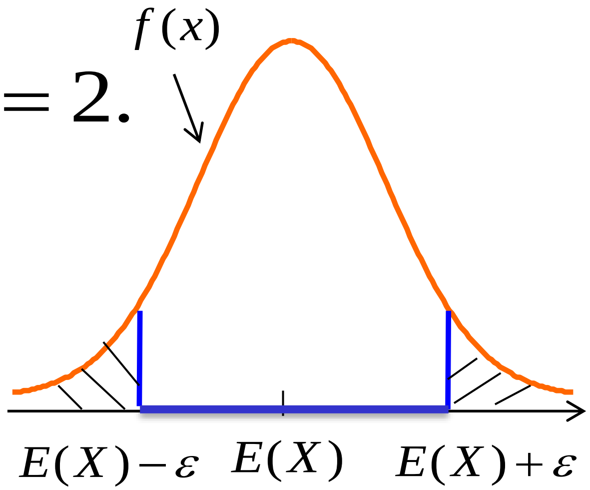
证明： 在定理5. 1. 1中

令$Y=X-E(X),k=2.$

则$E(|Y|^2)$

$=E(X-E(X))^2$

$=Var(X).$



例1.3　设$E(X)=\mu, Var(X)=\sigma^2$,

则 $P(|X-\mu|<3\sigma)\geq 1-\dfrac{\sigma^2}{(3\sigma)^2}=\dfrac{8}{9}$.

当$X\sim N(\mu,\sigma^2)$时,

$$P(|X-\mu|<3\sigma)=0.9974>\frac{8}{9}.$$

例1.4　设$X_1,\ldots,X_n$是随机变量序列,

$$E(X_n)=0, Var(X_n)=\frac{1}{n}, \text{则} X_n \xrightarrow{P} 0.$$

证明:利用切比雪夫不等式:$\forall \varepsilon > 0$,

$$0 \le P(|X_n - 0| \ge \varepsilon) \le \frac{Var(X_n)}{\varepsilon^2} = \frac{1}{n\varepsilon^2} \to 0.$$

$$\therefore \lim_{n\to\infty} P(|X_n - 0| \ge \varepsilon) = 0.$$

例1.5 在n重贝努里试验中，若已知每次试验

事件A出现的概率为0.75，

试利用切比雪夫不等式计算,

(1) 若n=7500, 估计A出现的频率在0.74至0.76

之间的概率至少有多大；

(2) 估计n, 使A出现的频率在0.74至0.76之

间的概率不小于0.90。

解：设在n重贝努里试验中，事件A出现的次数为X，

则$X \sim B(n, 0.75)$, $E(X) = np = 0.75n, Var(X) = npq = 0.1875n,$

又 $f_n(A) = \dfrac{X}{n}$, $P\left\{0.74 < \dfrac{X}{n} < 0.76\right\} = P\{0.74n < X < 0.76n\}$

$$= P\{|X - 0.75n| < 0.01n\}$$

$$\geq 1 - \frac{0.1875n}{(0.01n)^2} = 1 - \frac{1875}{n}$$

(1) $n = 7500$, $P\left\{0.74 < \dfrac{X}{n} < 0.76\right\} \geq 1 - \dfrac{1875}{7500} = 0.75$

(2) $P\left\{0.74 < \dfrac{X}{n} < 0.76\right\} \geq 1 - \dfrac{1875}{n} \geq 0.90 \quad \Rightarrow n \geq 18750.$

(三) 几个大数定律

定义5.1.2:

设$Y_1,\ldots,Y_n,\ldots$为一个随机变量序列,

若存在常数序列$\{c_n, n\geq 1\}$，使得

当$n\to+\infty$, $\dfrac{1}{n}\sum_{i=1}^{n}Y_i - c_n \xrightarrow{P} 0$,

即$\forall\varepsilon>0$，有$\lim\limits_{n\to\infty}P\{\left|\dfrac{1}{n}\sum_{i=1}^{n}Y_i - c_n\right|\geq\varepsilon\}=0$,

则称$\{Y_i, i\geq 1\}$服从(弱)大数定律.

特别地，当$c_n=c, n=1,2,\ldots$时，可写为

$$\frac{1}{n}\sum_{i=1}^{n}Y_i \xrightarrow{P} c, n\to+\infty.$$

定理（切比雪夫大数定律）：

设$X_1, X_2, \cdots, X_n, \cdots$相互独立，具有相同的数学期望$\mu$和相同的方差$\sigma^2$,

则当 $n \to +\infty$时, $\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{P} \mu.$

证明: $\because E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu,$

$$Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2} Var(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n},$$

利用切比雪夫不等式:

$$0 \le P(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| \ge \varepsilon) \le Var(\frac{1}{n}\sum_{i=1}^{n} X_i) \Big/ \varepsilon^2 = \frac{\sigma^2}{n\varepsilon^2} \to 0.$$

定理（辛钦大数定律）：

设$X_1, X_2, \cdots, X_n, \cdots$独立同分布，$EX_i = \mu,$

则当$n \to +\infty,$

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu.$$

推论：设$X_1, X_2, \cdots, X_n, \cdots$独立同分布，

若$h(x)$是连续函数，$E(h(X_1)) = a$存在，

则当$n \to +\infty$时，$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \xrightarrow{P} a.$

例1.6　设$X_1,\cdots,X_n,\cdots,$相互独立，$P\{X_i=0\}=1-\frac{1}{i},$

$$P\{X_i=\sqrt{i}\}=P\{X_i=-\sqrt{i}\}=\frac{1}{2i}。$$

试判断$\{X_i,i\geq 1\}$是否服从大数定律？

解：$E(X_i)=0,$

$$Var(X_i)=E(X_i^2)=0+(\sqrt{i})^2\cdot\frac{1}{2i}+(\sqrt{i})^2\cdot\frac{1}{2i}=1,$$

$$\therefore \frac{1}{n}\sum_{i=1}^{n}X_i\xrightarrow{P}0$$

例1.7：设$X_1,\cdots,X_n,\cdots$独立同分布，$X_1\sim U(-1,1)$.则

$$(1)\frac{1}{n}\sum_{k=1}^{n}X_k,\quad (2)\frac{1}{n}\sum_{k=1}^{n}|X_k|,\quad (3)\frac{1}{n}\sum_{k=1}^{n}X_k^2$$

依概率收敛于什么？

解：$E(X_1)=0 \quad \Rightarrow \frac{1}{n}\sum_{k=1}^{n}X_k\xrightarrow{P}0$

$$E(|X_1|)=\int_{-1}^{1}|x|\frac{1}{2}dx=\frac{1}{2}\Rightarrow\frac{1}{n}\sum_{k=1}^{n}|X_k|\xrightarrow{P}\frac{1}{2},$$

$$E(X_1^2)=\int_{-1}^{1}x^2\frac{1}{2}dx=\frac{1}{3}\Rightarrow\frac{1}{n}\sum_{k=1}^{n}X_k^2\xrightarrow{P}\frac{1}{3}.$$

例1.8 设$X_1,\cdots,X_n,\cdots$独立同分布，$X_1 \sim U(0,1)$,

则$\sqrt[n]{X_1 X_2 \ldots X_n}$依概率收敛吗？

如果依概率收敛，收敛于什么？

解：令$Y_n = \sqrt[n]{X_1 \ldots X_n}$， $Z_n = \ln Y_n$

则$Z_n = \dfrac{1}{n}(\ln X_1 + \ldots + \ln X_n)$,

$$E(\ln X_1) = \int_0^1 \ln x dx$$

$$= x\ln x\Big|_0^1 - \int_0^1 dx = -1,$$

由辛钦大数定律的推论，$Z_n \xrightarrow{P} -1.$

由依概率收敛的性质，$Y_n = e^{Z_n} \xrightarrow{P} e^{-1}.$

定理（贝努里大数定律）：

设n_A为n重贝努里试验中事件A发生的次数，并记事件A在每次试验中发生的概率为p，则有

$$\frac{n_A}{n} \xrightarrow{P} p, \text{当} n \to +\infty \text{时}.$$

证明思路：易见$n_A = \sum_{i=1}^{n} X_i$, 其中

$$X_i = \begin{cases} 1, & \text{第}i\text{次试验中}A\text{发生;} \\ 0, & \text{第}i\text{次试验中}A\text{不发生;} \end{cases} \qquad i = 1, 2, \cdots, n,$$

$X_i \sim B(1, p)$, 且相互独立.

(或直接用切比雪夫不等式证明).

大数定律的重要意义：

贝努里大数定律建立了在大量重复独立试验中事件出现频率的稳定性，正因为这种稳定性，概率的概念才有客观意义.

贝努里大数定律还提供了通过试验来确定事件概率的方法，既然频率n_A/n与概率p有较大偏差的可能性很小，因此可以通过做试验确定某事件发生的频率并把它作为相应的概率估计，这是一种参数估计法，该方法的重要理论基础之一就是大数定律.

5.2 中心极限定理

定理5.2.1(独立同分布的中心极限定理)：

设$X_1, X_2, \cdots, X_n, \cdots$独立同分布，

$E\left(X_i\right)=\mu, Var\left(X_i\right)=\sigma^2$,则对任意实数$x$,

$$\lim_{n\to\infty} P\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \le x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \Phi(x).$$

因此当n充分大时 $\sum_{i=1}^{n} X_i \overset{近似}{\sim} N(n\mu, n\sigma^2).$

$$\frac{1}{n}\sum_{i=1}^{n} X_i \overset{近似}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

推论(棣莫弗—拉普拉斯中心极限定理):

设n_A为n重贝努里试验中A发生的次数,

$P(A)=p(0<p<1)$,则对任何实数x,有:

$$\lim_{n\to+\infty} P\left(\frac{n_A-np}{\sqrt{np(1-p)}}\le x\right)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt=\Phi(x).$$

即,$B(n,p)\overset{近似}{\sim} N(np,np(1-p))$,当$n$充分大时.

证明:令$X_i=\begin{cases}1, & 第i次试验A发生,\\ 0, & 第i次试验A不发生,\end{cases}\quad i=1,2,\cdots,n.$

$X_1,X_2,\ldots X_n$独立同$\sim B(1,p)$. $\quad n_A=\sum_{i=1}^{n}X_i$,由定理可得.

例2.1 某宴会上提供一瓶6升(l)的大瓶法国红酒,假定与会者每次所倒红酒的重量服从同一分布,期望值为100毫升(ml),标准差为32毫升.若每次所倒红酒都是相互独立的,试问:倒了55次后该瓶红酒仍有剩余的概率约为多少?

解：设X_i为第i次所倒的红酒量(单位:ml)，则

X_i独立同分布，$E(X_i)=100, Var(X_i)=32^2, i=1,\cdots,55.$

根据独立同分布的中心极限定理知，

$$\frac{\sum_{i=1}^{55} X_i - 55\times 100}{32\sqrt{55}} \overset{近似}{\sim} N(0,1),$$

所以，所求概率为 $P\{\sum_{i=1}^{55} X_i < 6000\}$

$$=P\{\frac{\sum_{i=1}^{55} X_i - 55\times 100}{32\sqrt{55}} < \frac{6000-55\times 100}{32\sqrt{55}}\} \approx \Phi\left(\frac{6000-55\times 100}{32\sqrt{55}}\right)$$

$$=\Phi(2.11)=0.9826.$$

例2.2 某校1500名学生选修“概率论与数理统计”课程,共有10名教师主讲此课,假定每位学生可以随意地选择一位教师(即,选择任意一位教师的可能性均为1/10),而且学生之间的选择是相互独立的.问:每位教师的上课教室应该设有多少座位才能保证该班因没有座位而使学生离开的概率小于5%.

解：由于每位学生可以随意地选择一位老师，因此我们只需要考虑教师甲的上课教室座位数即可. 令Y表示选择教师甲的学生人数，则

$$Y \sim B(1500, 1/10).$$

根据棣莫弗–拉普拉斯中心极限定理知，

$$\frac{Y - 1500 \times \frac{1}{10}}{\sqrt{1500 \times \frac{1}{10} \times \frac{9}{10}}} \overset{\text{近似}}{\sim} N(0,1).$$

设教室座位数为a, 则由题意

$$95\% \le P\{Y \le a\} \approx \Phi\left(\frac{a-1500\times\frac{1}{10}}{\sqrt{1500\times\frac{1}{10}\times\frac{9}{10}}}\right)$$

查表得 $\Phi(1.645)=95\%$, 故需 $\dfrac{a-150}{\sqrt{135}} \ge 1.645,$

解得 $a \ge 169.11.$

故每位老师的上课教室应该至少设170个座位才能保证因座位不够使学生离开的概率小于5%.

例2.3　设随机变量$X_1, X_2, \cdots, X_{50}$,相互独立同分布，$X_1 \sim U(-1,1)$.　分别求下列随机变量的近似分布.

(1) $\frac{1}{50}\sum_{k=1}^{50} X_k$,　(2) $\frac{1}{50}\sum_{k=1}^{50} |X_k|$,　(3) $\frac{1}{50}\sum_{k=1}^{50} X_k^2$.

解：由独立同分布的中心极限定理，

$$\frac{1}{50}\sum_{k=1}^{50}X_k,\quad \frac{1}{50}\sum_{k=1}^{50}|X_k|,\quad \frac{1}{50}\sum_{k=1}^{50}X_k^2$$均近似服从正态分布。

因为，$E(X_1)=0, Var(X_1)=\frac{4}{12}=\frac{1}{3}, \Rightarrow \frac{1}{50}\sum_{k=1}^{50}X_k \overset{近似}{\sim} N(0,\frac{1}{150}),$

$$E(|X_1|)=\frac{1}{2}, Var(|X_1|)=E(X_1^2)-[E(|X_1|)]^2=\frac{1}{12},$$

$$\Rightarrow \frac{1}{50}\sum_{k=1}^{50}|X_k| \overset{近似}{\sim} N(\frac{1}{2},\frac{1}{600}),$$

$$E(X_1^2)=\frac{1}{3},\quad Var(X_1^2)=E(X_1^4)-[E(X_1^2)]^2=\frac{1}{5}-\frac{1}{9}=\frac{4}{45},$$

$$\Rightarrow \frac{1}{50}\sum_{k=1}^{50}X_k^2 \overset{近似}{\sim} N(\frac{1}{3},\frac{2}{1125}).$$

例2.4　在n重贝努里试验中，若已知每次试验事件A出现的概率为0.75，试利用中心极限定理计算,

(1)若n=7500,估计A出现的频率在0.74至0.76之间的概率近似值；

(2)估计n,使A出现的频率在0.74至0.76之间的概率不小于0.90。

解：设事件***A***出现的次数为***X***,则 $X \sim B\left(n, 0.75\right)$,

$$E\left(X\right) = 0.75n, \quad Var\left(X\right) = 0.1875n,$$

$$P\left\{0.74 < \frac{X}{n} < 0.76\right\}$$

$$\approx \Phi\left(\frac{0.76n - 0.75n}{\sqrt{0.1875n}}\right) - \Phi\left(\frac{0.74n - 0.75n}{\sqrt{0.1875n}}\right) = 2\Phi\left(\frac{\sqrt{3n}}{75}\right) - 1$$

(1) $n = 7500,\ P\left\{0.74 < \frac{X}{n} < 0.76\right\} \approx 2\Phi(2) - 1 = 0.9544$

(2)由$2\Phi\left(\frac{\sqrt{3n}}{75}\right) - 1 \geq 0.9,\ \Rightarrow \Phi\left(\frac{\sqrt{3n}}{75}\right) \geq 0.95,$

$$\frac{\sqrt{3n}}{75} \geq 1.645, \quad n \geq 5074.$$

切比雪夫不等式估计$n \geq 18750.$